Computer Vision: Fall 2022 — Lecture 16

Dr. Karthik Mohan

Univ. of Washington, Seattle

November 30, 2022

References

Generic ML/DL

1. Good Book for Machine Learning Concepts
2. Deep Learning Reference

CNN

1. Convolutional Neural Networks for Visual Recognition
2. Convolutional Neural Net Tutorial
3. CNN Transfer Learning
4. PyTorch Transfer Learning Tutorial

CNN Publication References

CNN surveys

1. Convolutional Neural Networks: A comprehensive survey, 2019
2. A survey of Convolutional Neural Networks: Analysis, Applications, and Prospects, 2021

CNN Archs

1. GoogLeNet
2. Top models on ImageNet
3. ResNet ILSVRC paper

Object Detection and Image Segmentation References

Object Detection

1. A survey of modern deep learning based object detection methods
2. YOLO Survey
3. YOLO Original Paper

Image Captioning

Comprehensive

1. From Show to Tell: A survey on Deep Learning-based Image Captioning
2. A survey of image captioning models

Lightweight

Today

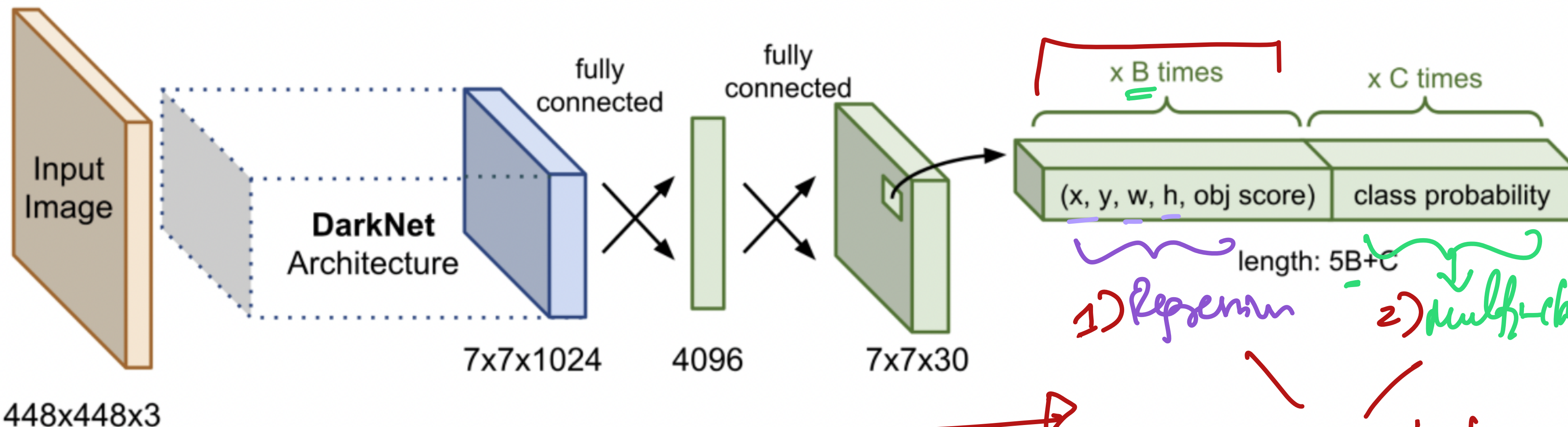
1. Recap of YOLO for object detection
2. Image Captioning intro and models

YOLO - Single Stage Detection

You only Look once!

Image

1) multiple Bounding Boxes for each Grid cell
2) class



1) Regression

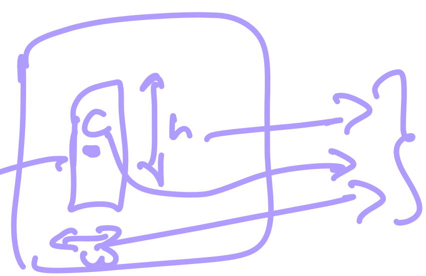
2) multi-class

Combine the two

Yolo v1
v2
v3
v4
v5

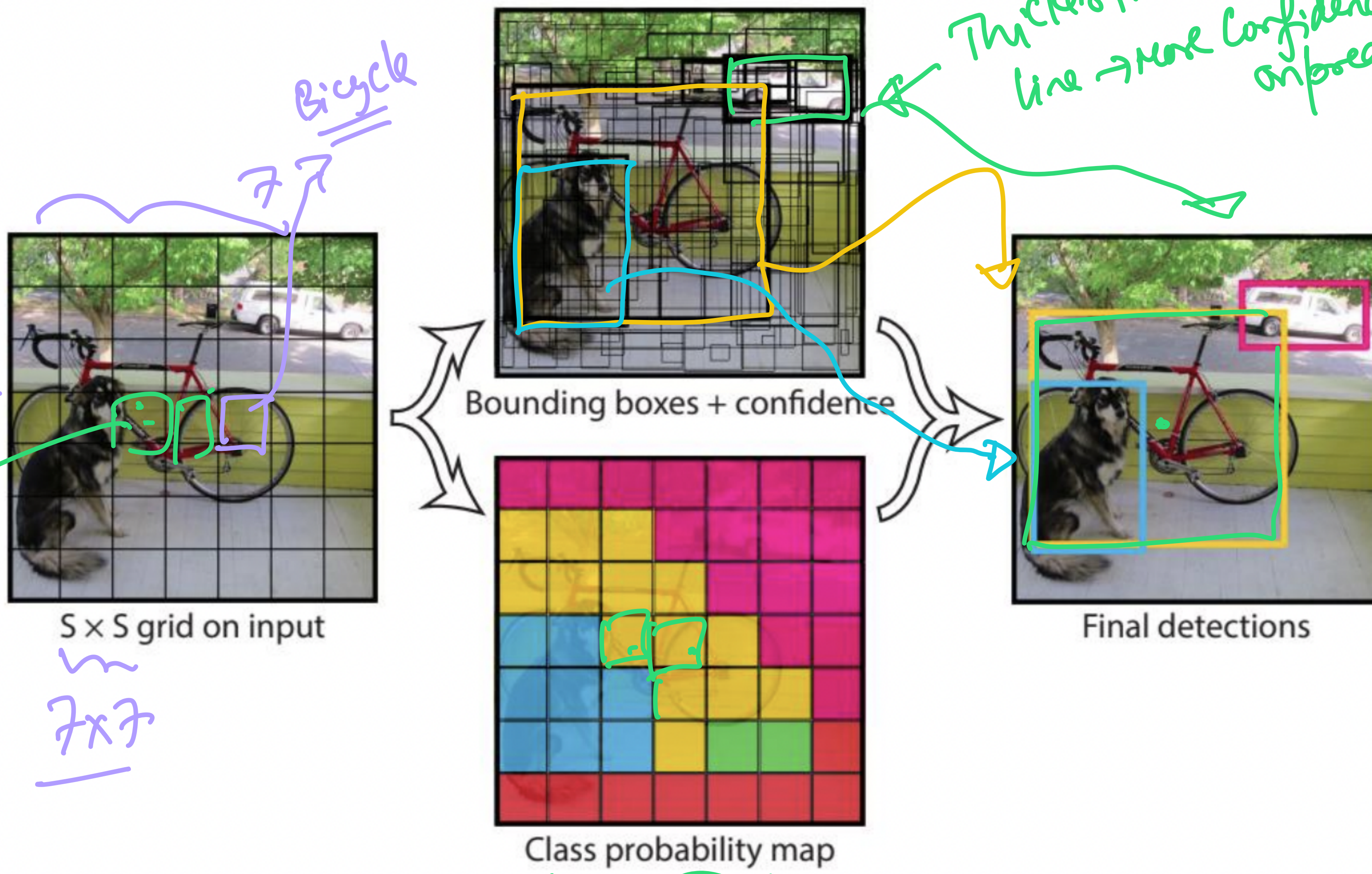
Improvement

obj ctr



continuous valued hence Regression!

YOLO Breakdown



Bicycle

7

7

Center

7x7

Thicker the line → more confidence on prediction!

YOLO benefits

1. **Single pass**

YOLO benefits

1. **Single pass**
2. **Fast**

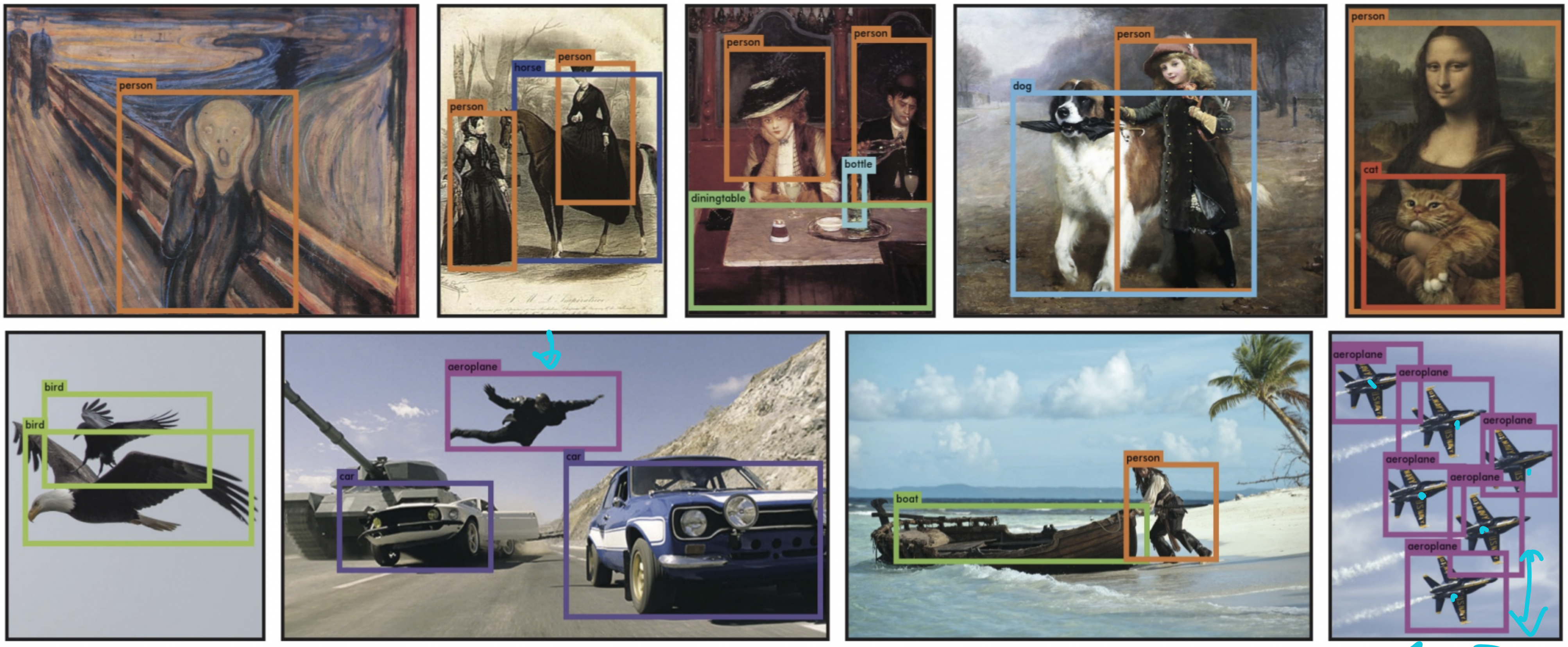
YOLO benefits

1. **Single pass**
2. **Fast**
3. **Global reasoning**

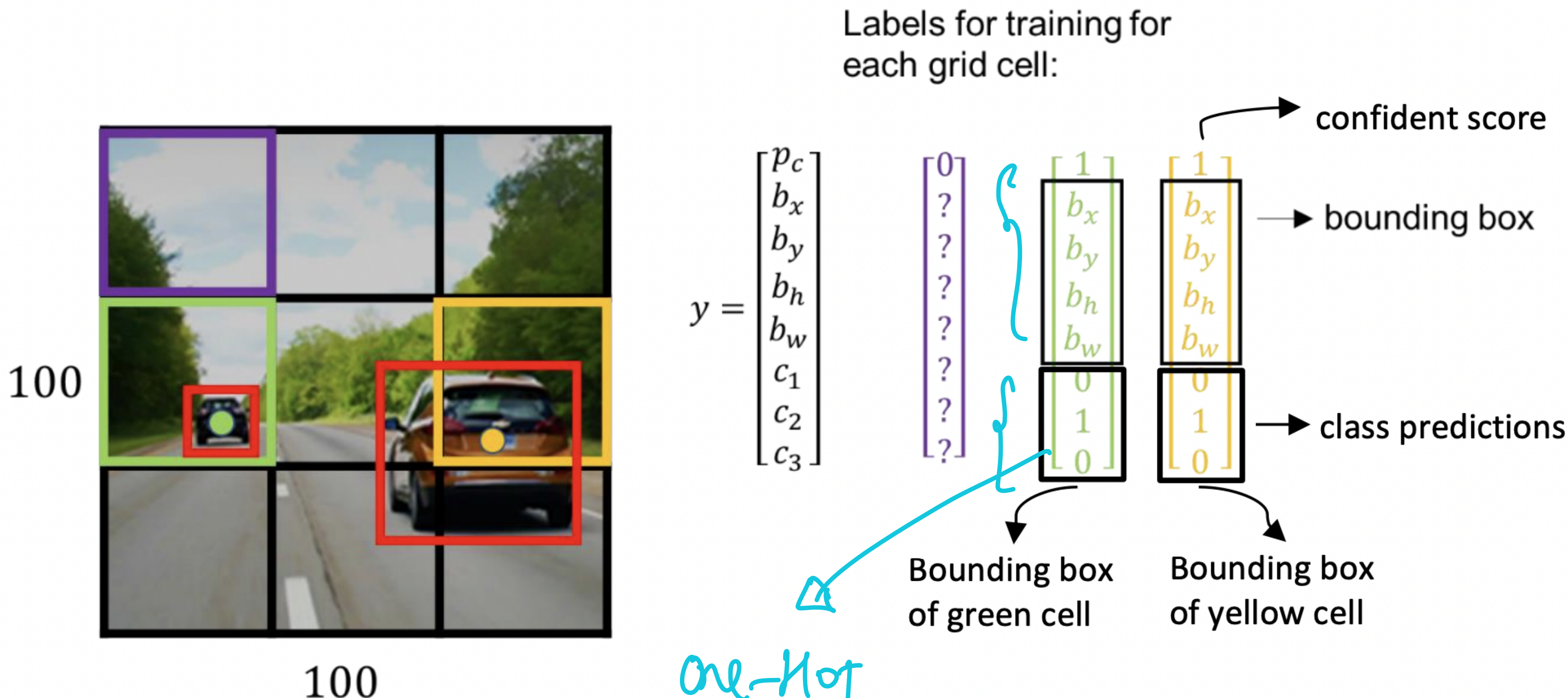
YOLO benefits

1. **Single pass**
2. **Fast**
3. **Global reasoning**
4. **More generalized representations**

YOLO examples



YOLO labels

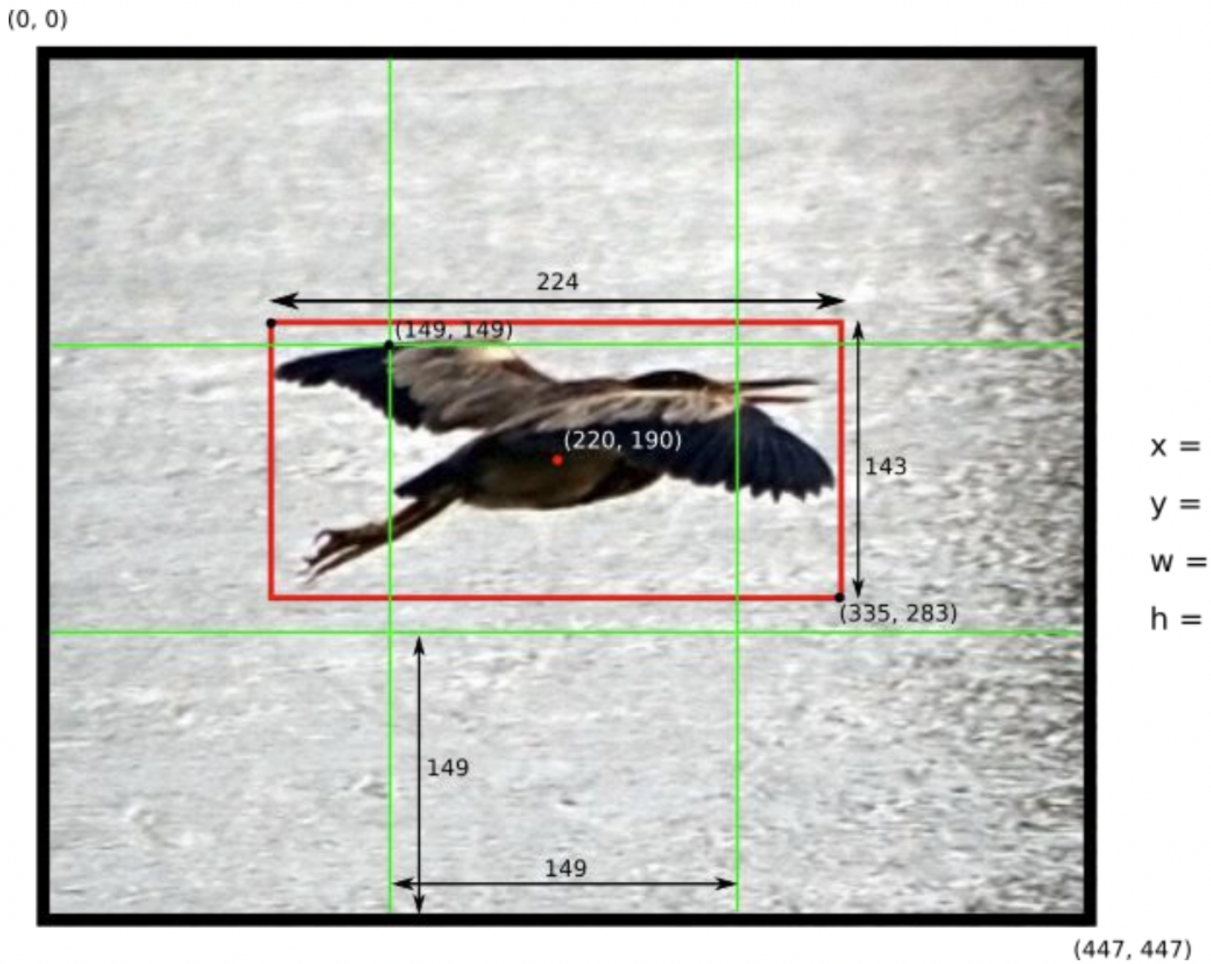


One-Hot Encoding

$S \times S \times (S \times B + C)$

#Labels for an Image

YOLO Bounding Box



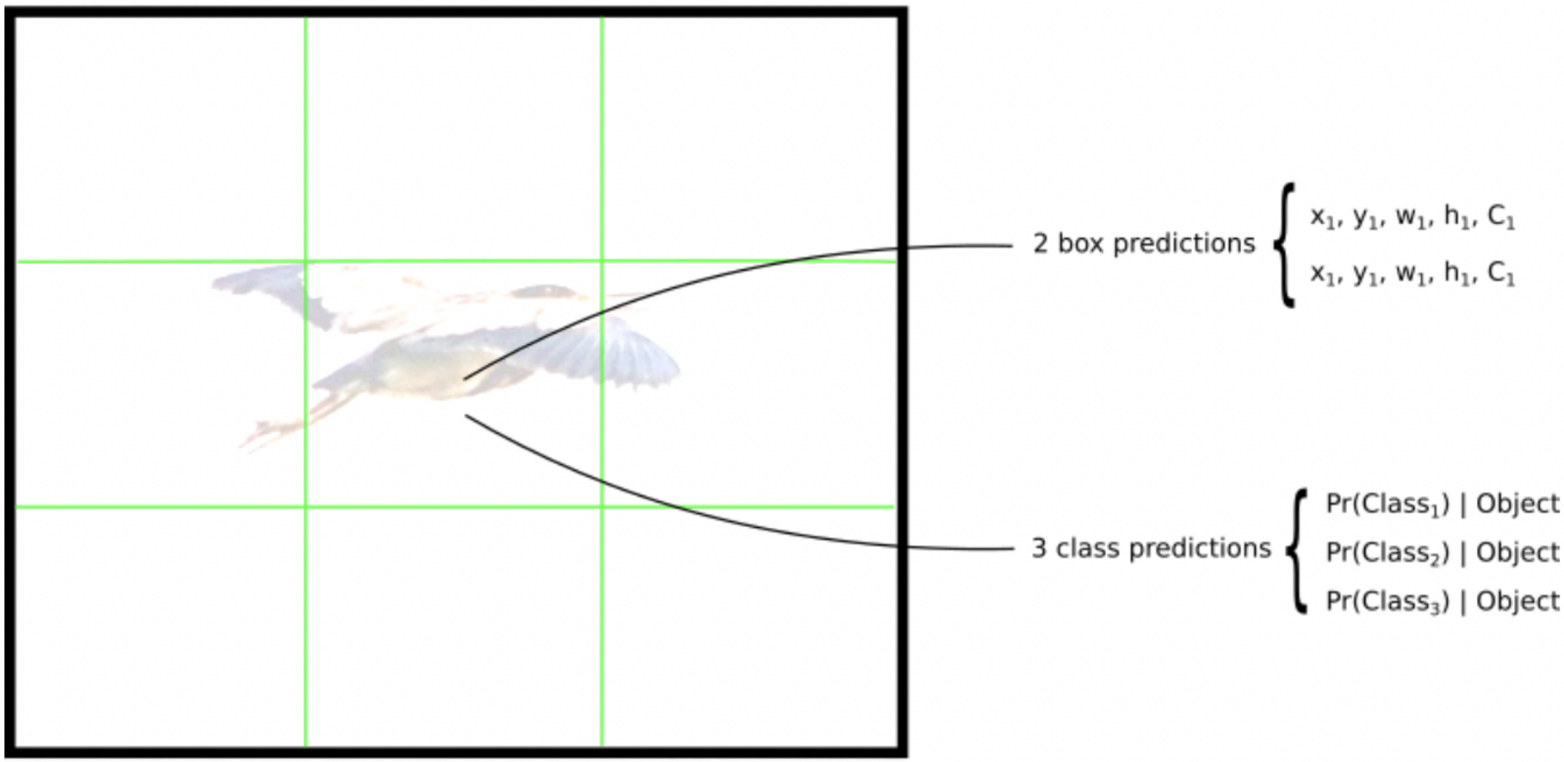
x = (220-149) / 149 = 0.48

y = (190-149) / 149 = 0.28

w = 224 / 448 = 0.50

h = 143 / 448 = 0.32

YOLO Bounding Box 2



YOLO and Regression

YOLO Loss Function - Regression!

YOLO loss function turns out to be just like a Regression Loss! Why Regression?

Body weight

Height → θ weight

Linear Regression Model:-

$h = wx\,\theta + b$

bodyweight

parameter

bias

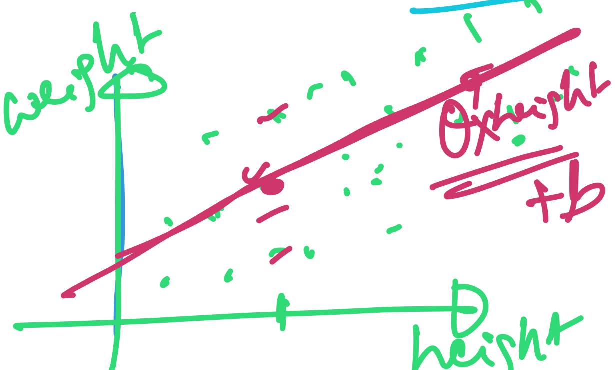
parameters

$\min_{\theta} \|X\theta - y\|_2^2$

Input

param

output



YOLO

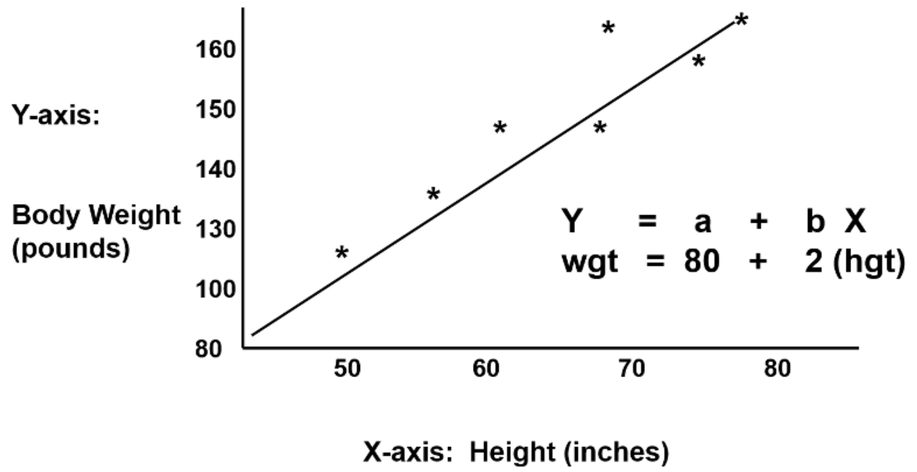
Image | Gridcells → Bounding Box + classes

YOLO and Regression

YOLO Loss Function - Regression!

YOLO loss function turns out to be just like a Regression Loss! Why Regression?

Linear Regression Classic Example



ICE #1

$e_1 = +0.1$ $e_2 = -0.2$ $e_3 = +0.1$

Sharpness (pred, truth)

(0.5, 0.4), (0.6, 0.8), (0.3, 0.2)

Regression (2 mins)

You want to predict the 'sharpness' of an image when the input is an image. Sharpness for this exercise is defined on a continuous scale between 0 and 1. The training data looks like $\{Image, Sharpness\}$ where Image is the input and Sharpness (on a continuous scale) is the output. You devise an ingenious loss function as follows: Take the prediction $\hat{y}_i$ of the sharpness, subtracts it from the ground truth sharpness y_i, and obtain the error, e_i. Define the loss, $L = \frac{1}{N}\sum_i e_i$. You then minimize the loss as you hope a good model for sharpness would give zero errors and hence a close to zero loss. Optimizing the loss function:

N - # examples

1. Will help you train a good model for sharpness
2. Is a good idea but may have to watch out for overfitting
3. Would not be a good idea
4. Could result in a model with overall zero error but poor individual predictions

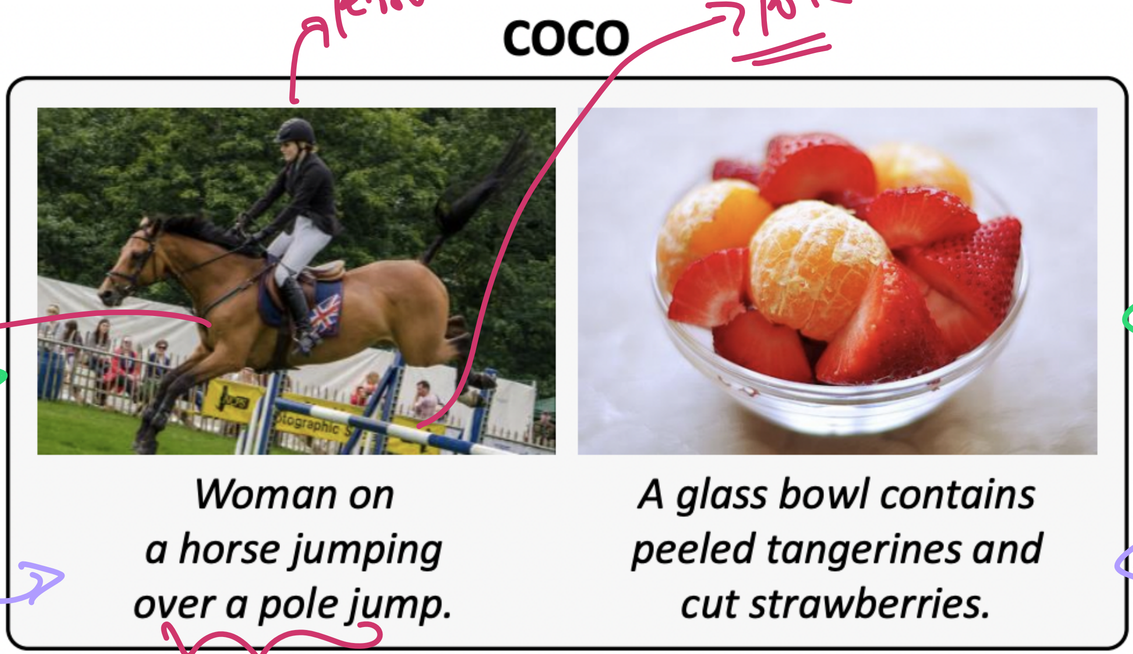
Better

$L = \frac{1}{N}\sum_i e_i^2$

Next Topic: Image Captioning Models

COCO Data Set

MS-COCO



Person

Pole

Horse

Image

Image

Caption

Caption

NLP task

COCO Data Set

COCO

Woman on a horse jumping over a pole jump.

A glass bowl contains peeled tangerines and cut strawberries.

COCO

Frequency
word cloud

Why Image Captioning?

1. Virtual Assistants

Why Image Captioning?

1. Virtual Assistants
2. Visually impaired assistance

Why Image Captioning?

1. Virtual Assistants
2. Visually impaired assistance
3. Robotics

Why Image Captioning?

1. Virtual Assistants
2. Visually impaired assistance
3. Robotics
4. Any other use case?

CUB-200 Data Set

CUB-200

1) What's the right level of detail to describe an image?

2) Style(Net)

↳ Style in which captions generated

— Humorous . . .

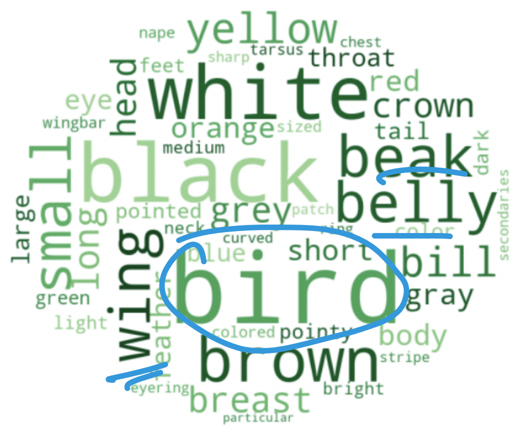
— Serious

CUB-200 Data Set

CUB-200

This bird is blue with white on its chest and has a very short beak.

CUB-200



Fashion Data Set

(Product)

Fashion Captioning

A decorative leather padlock on a compact bag with croc embossed leather.

Fashion Data Set

Fashion Captioning

A decorative leather padlock on a compact bag with croc embossed leather.

Fashion Captioning

Text Caps Data Set

TextCaps

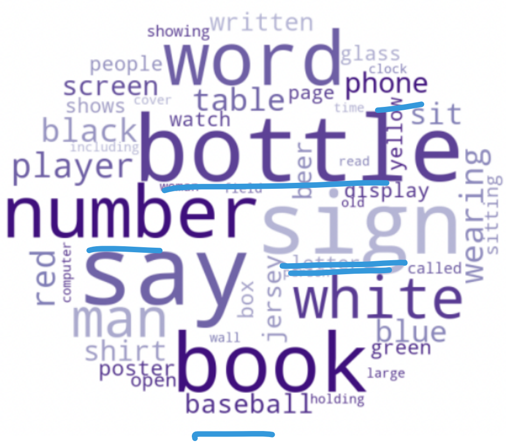
The billboard displays 'Welcome to Yakima The Palm Springs of Washington'.

Text Caps Data Set

TextCaps

The billboard displays 'Welcome to Yakima The Palm Springs of Washington'.

TextCaps

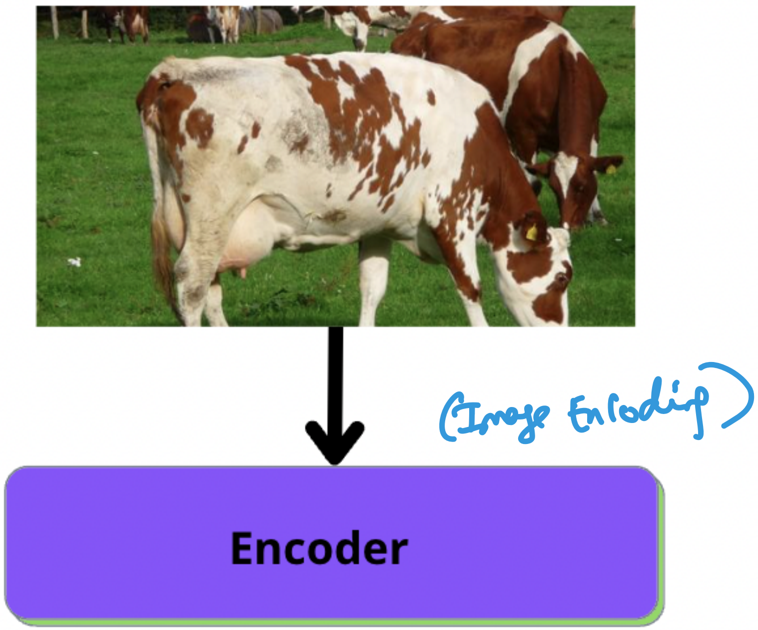


Encoder-Decoder Model for Image Captioning

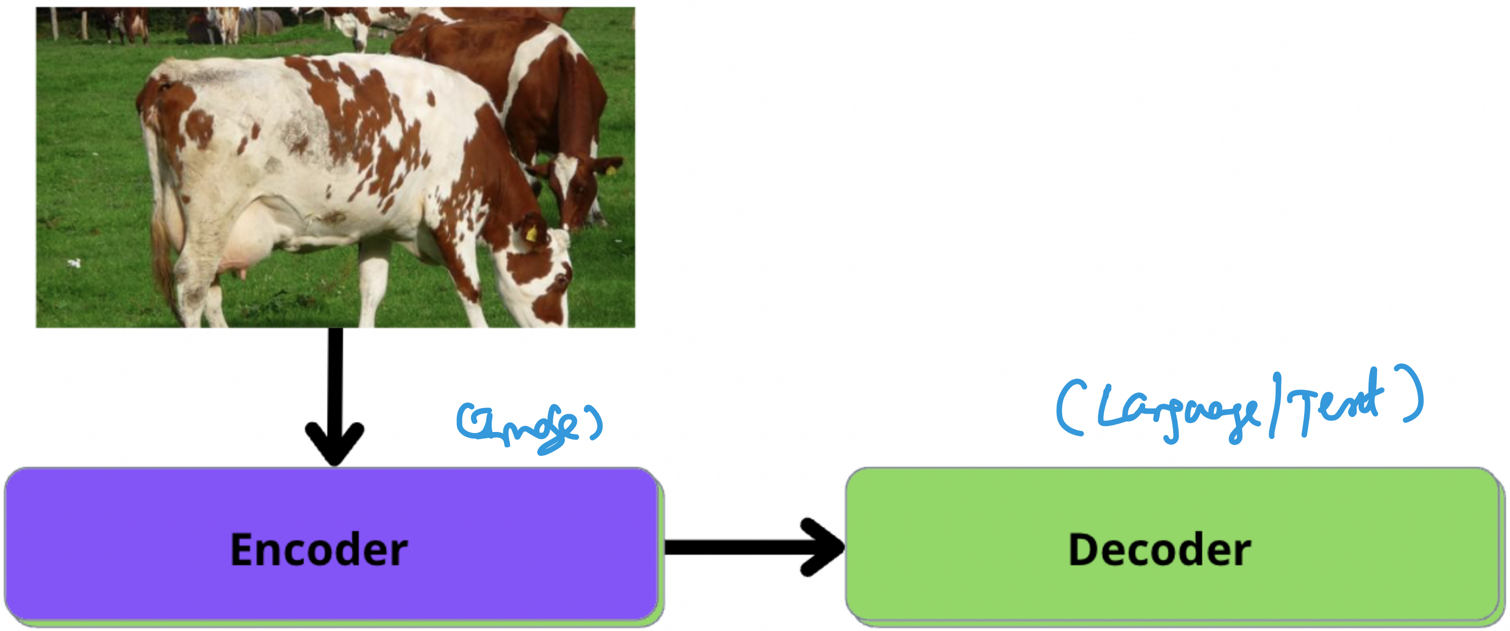
Encoder-Decoder Model for Image Captioning

← Image

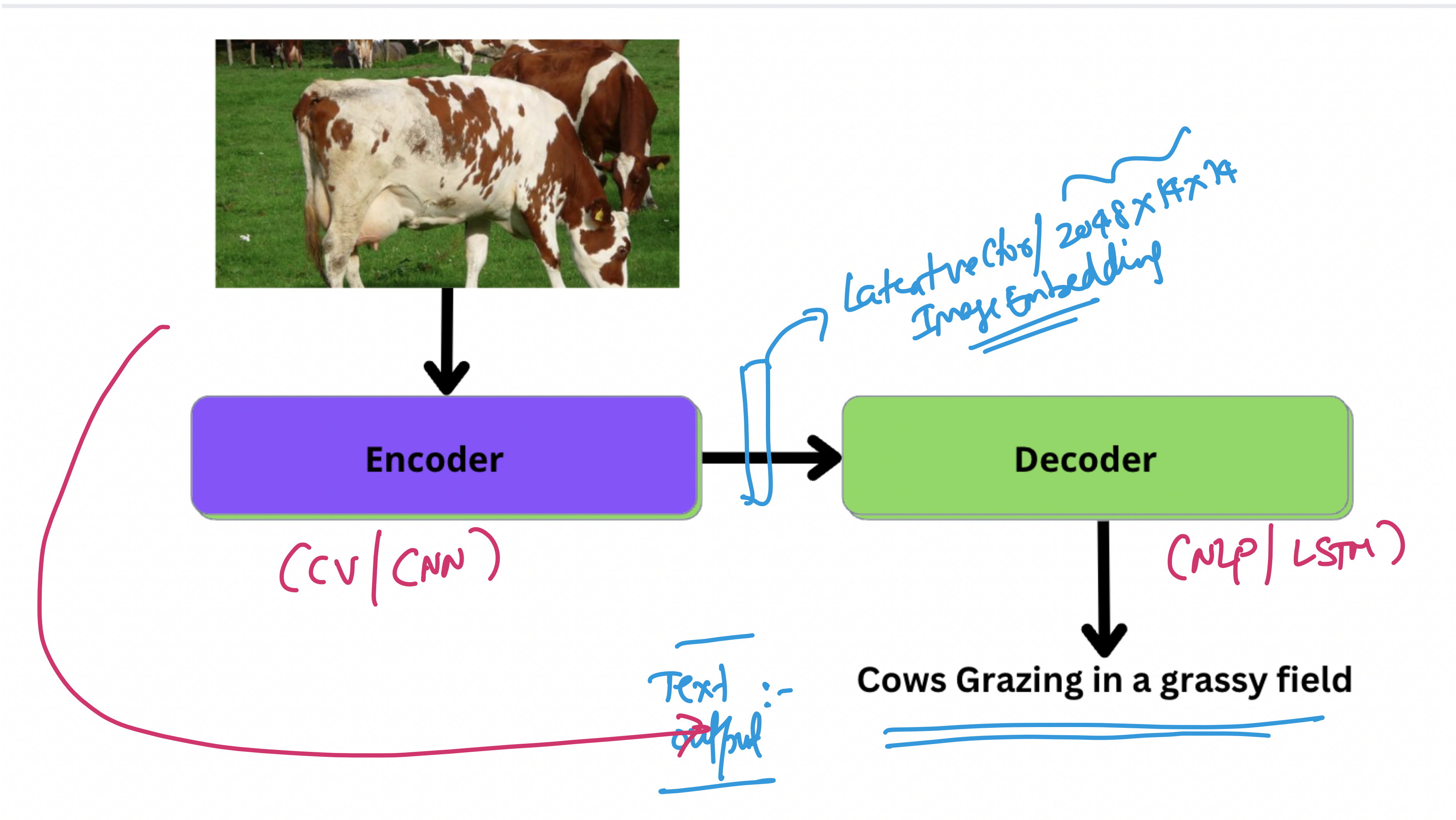
Encoder-Decoder Model for Image Captioning



Encoder-Decoder Model for Image Captioning



Encoder-Decoder Model for Image Captioning



Encoder-Decoder Model for Image Captioning

Image Encoder
(Typically CNN)

Language Decoder
(Typically RNN/LSTM)

Cows Grazing in a grassy field

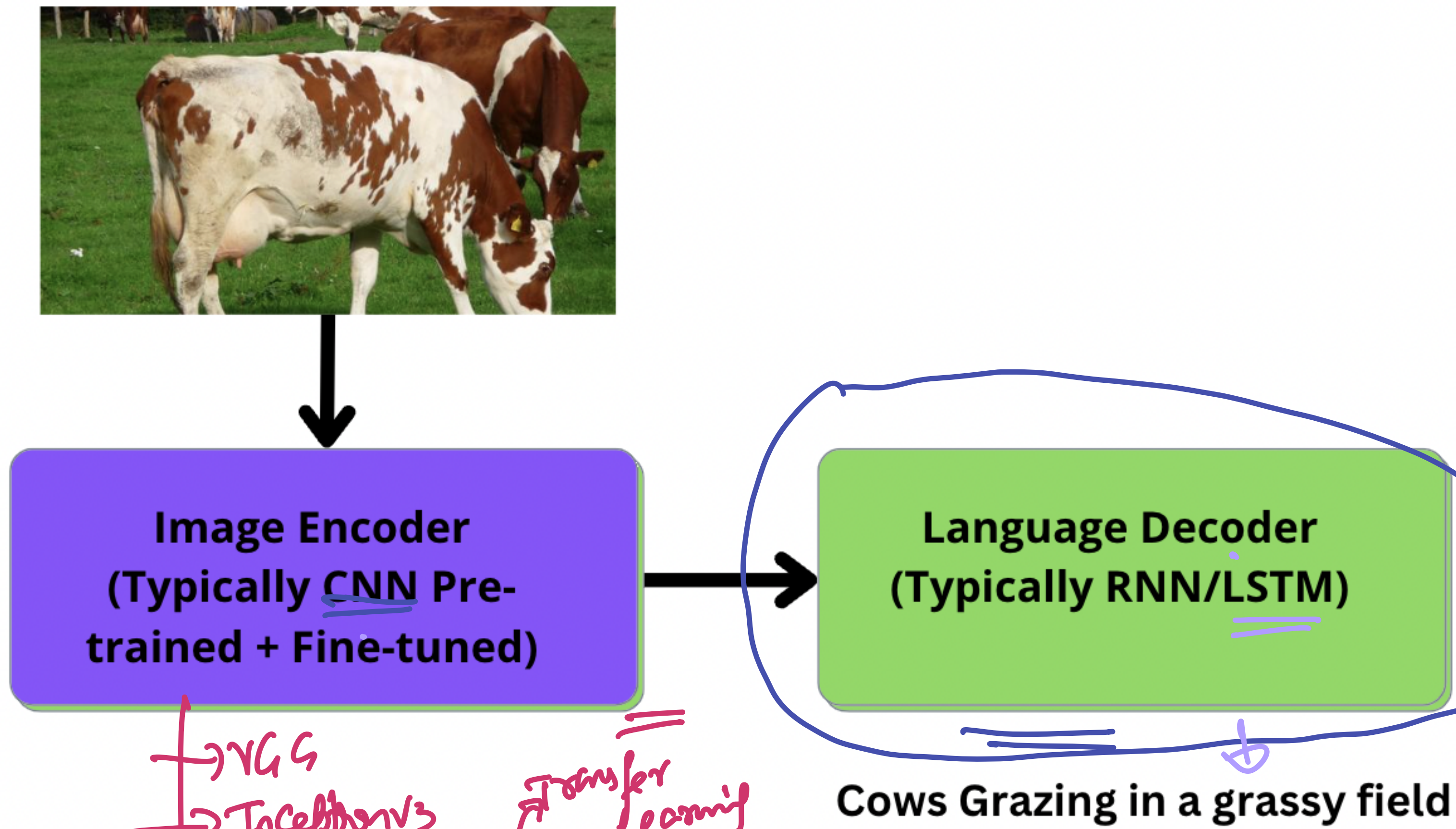
Region CNN

CNN R-CNN

RNN

Recurrent Neural Networks

Encoder-Decoder Model for Image Captioning

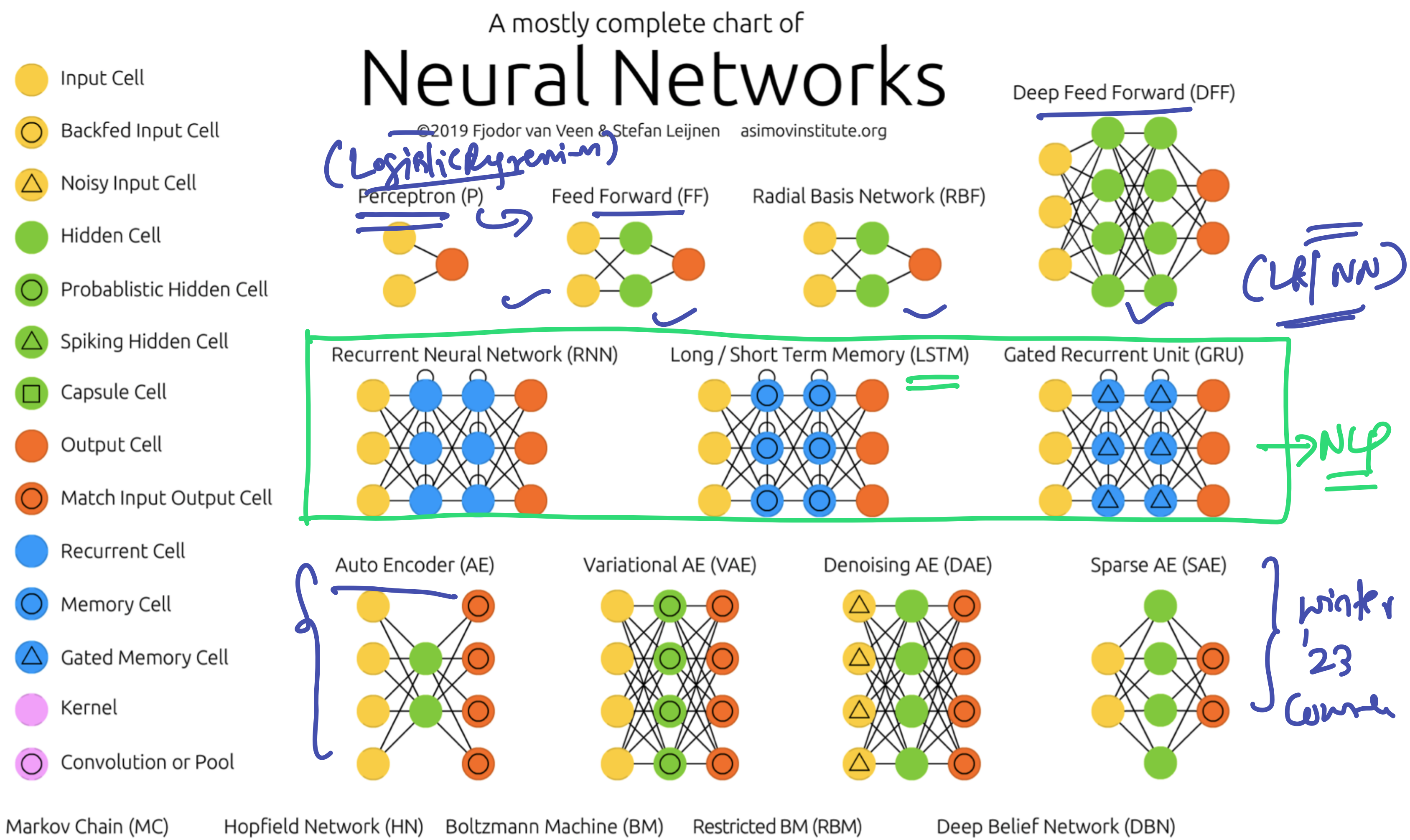


→ VGG
→ InceptionV3
→ ResNet (Pre-Trained + Fine-Tuning)

Transfer Learning

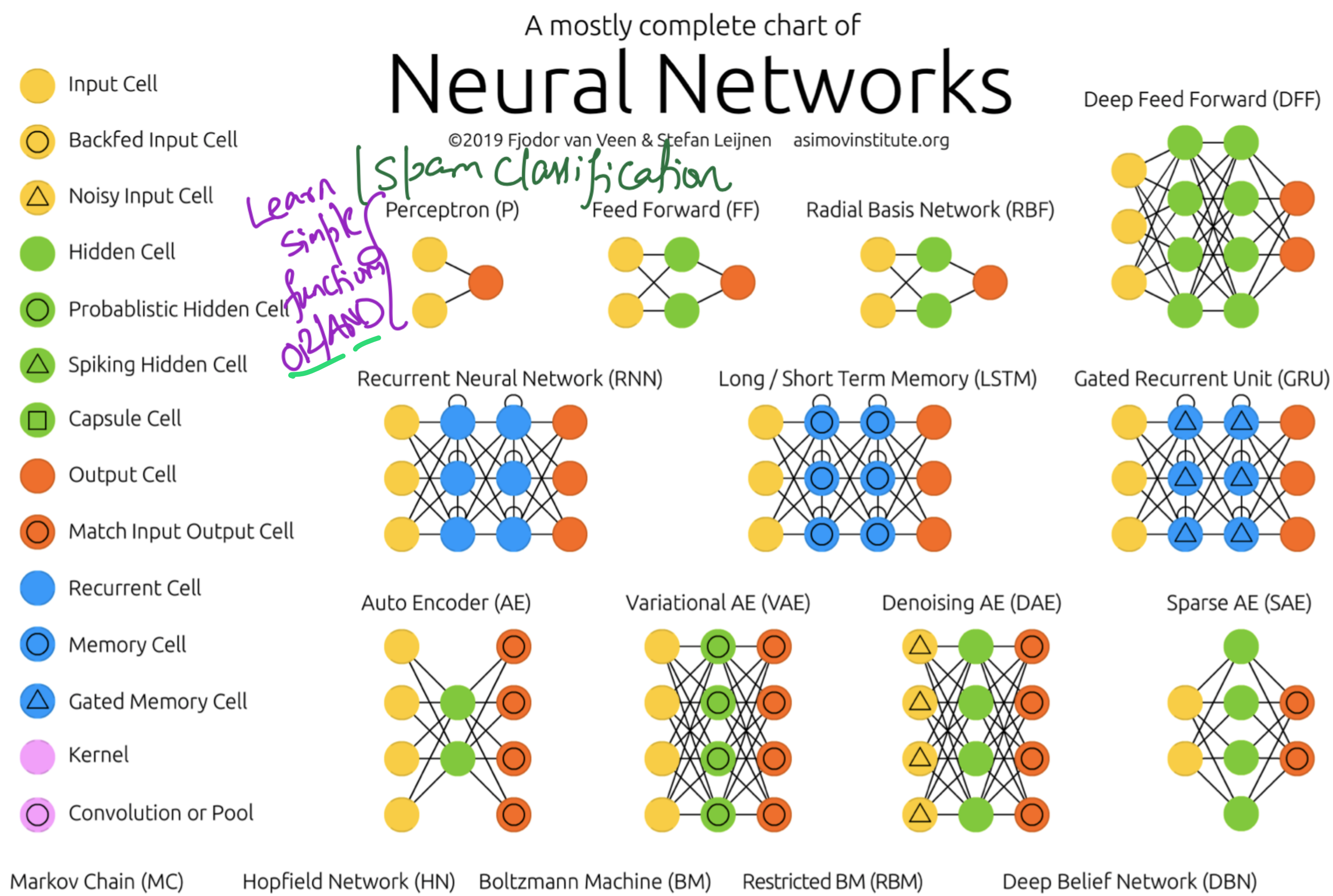
Deep Learning Zoo

Neural Networks Zoo



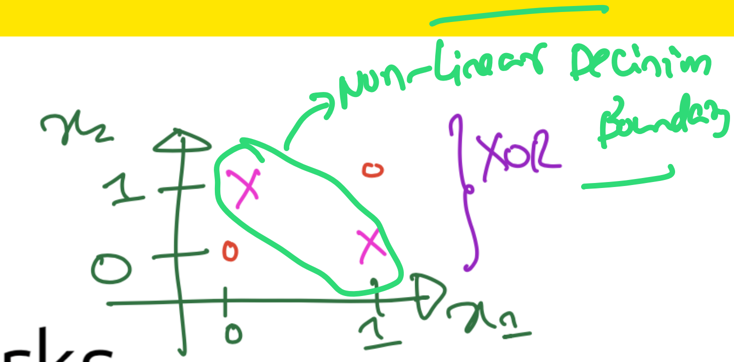
Deep Learning Zoo

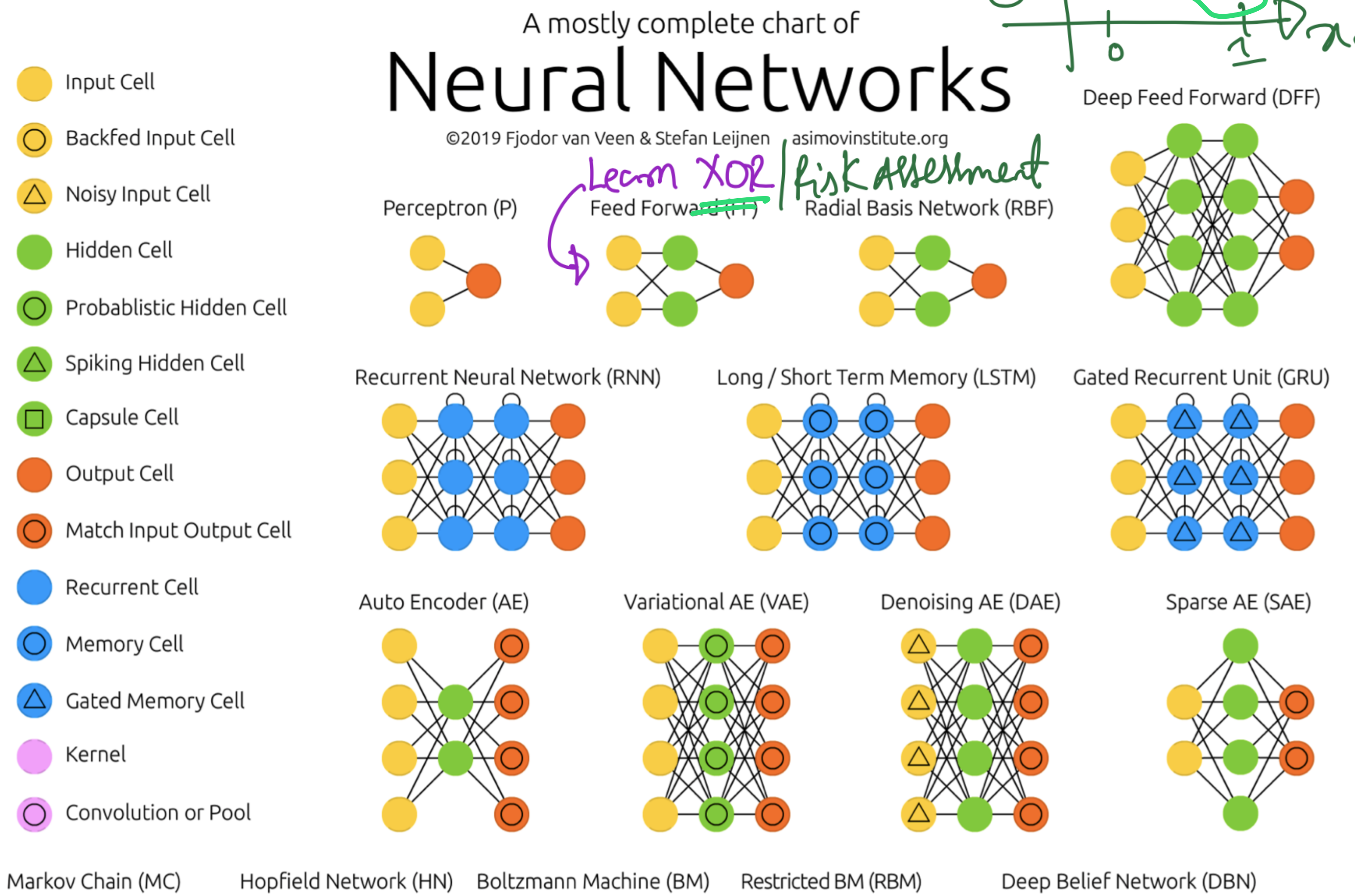
Neural Networks Zoo



Deep Learning Zoo

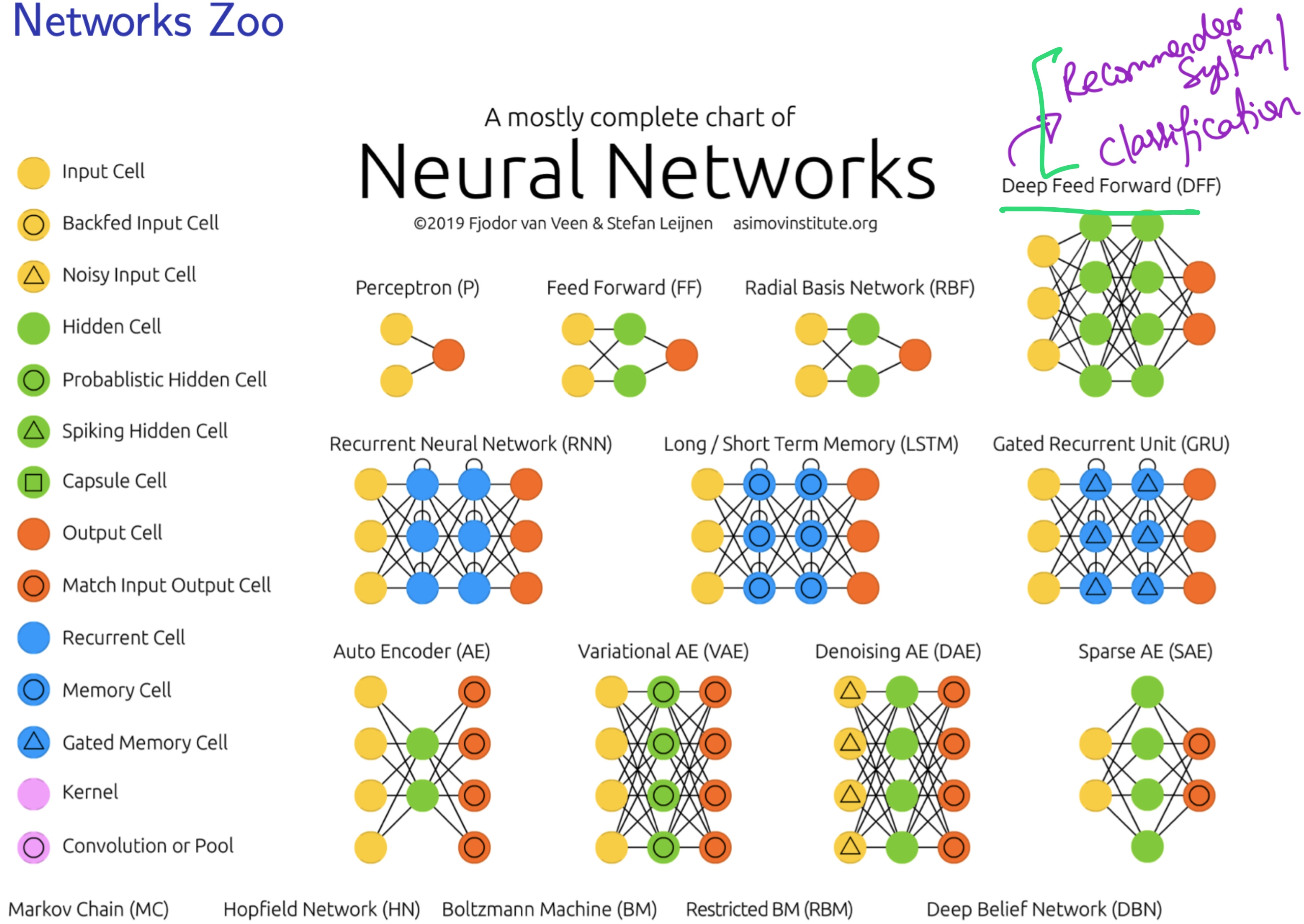
Neural Networks Zoo





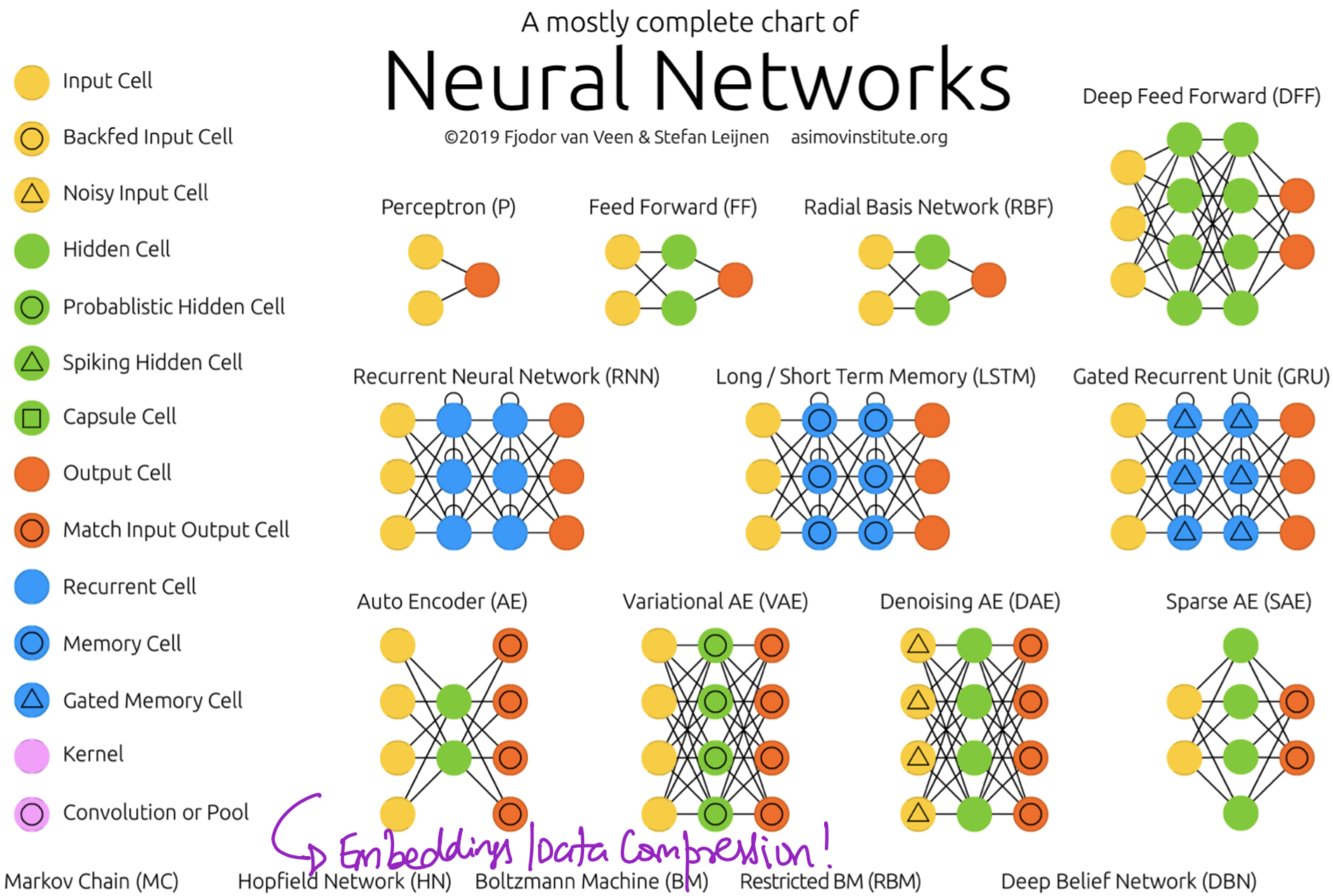
Deep Learning Zoo

Neural Networks Zoo



Deep Learning Zoo

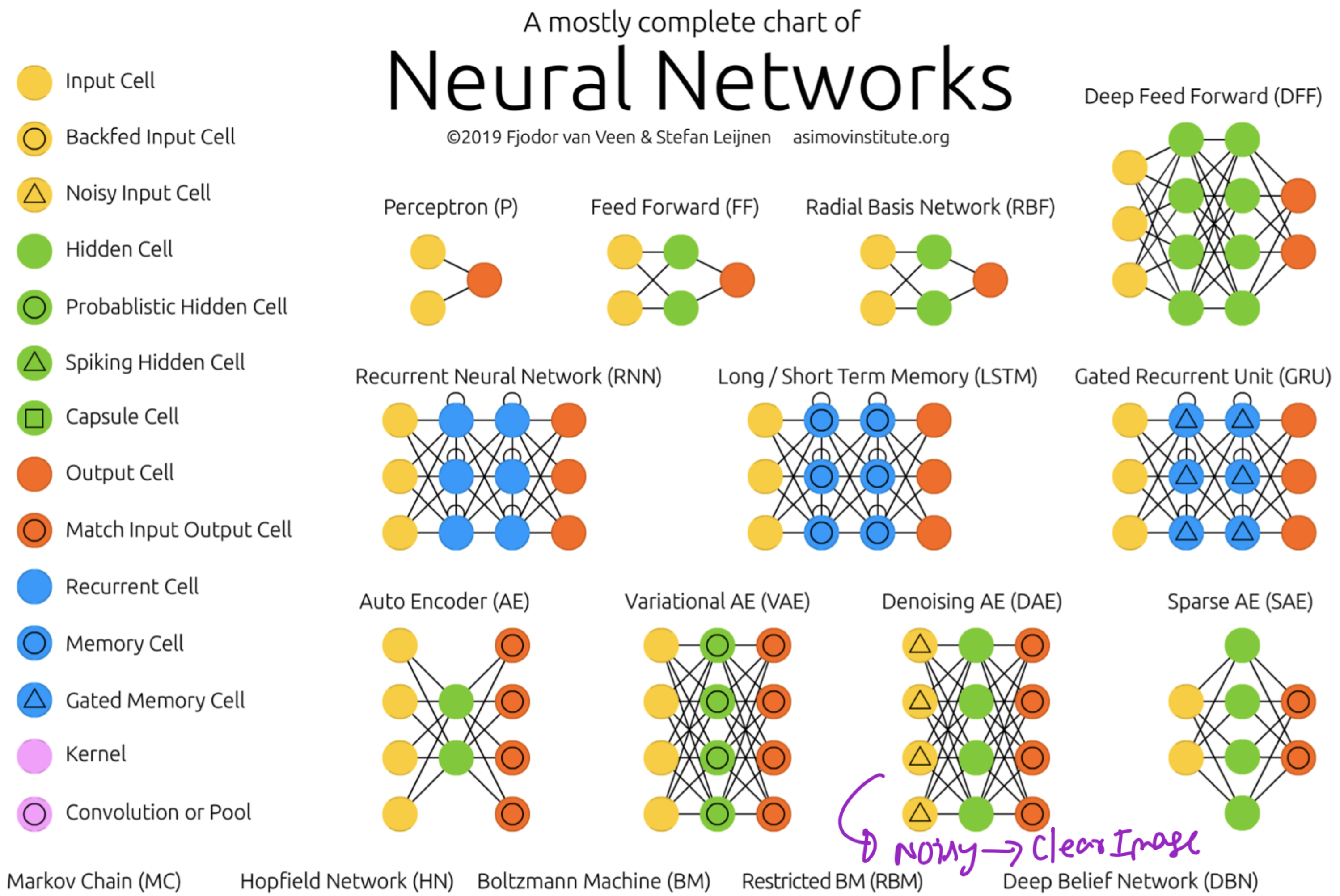
Neural Networks Zoo



→ Embeddings / Data Compression!

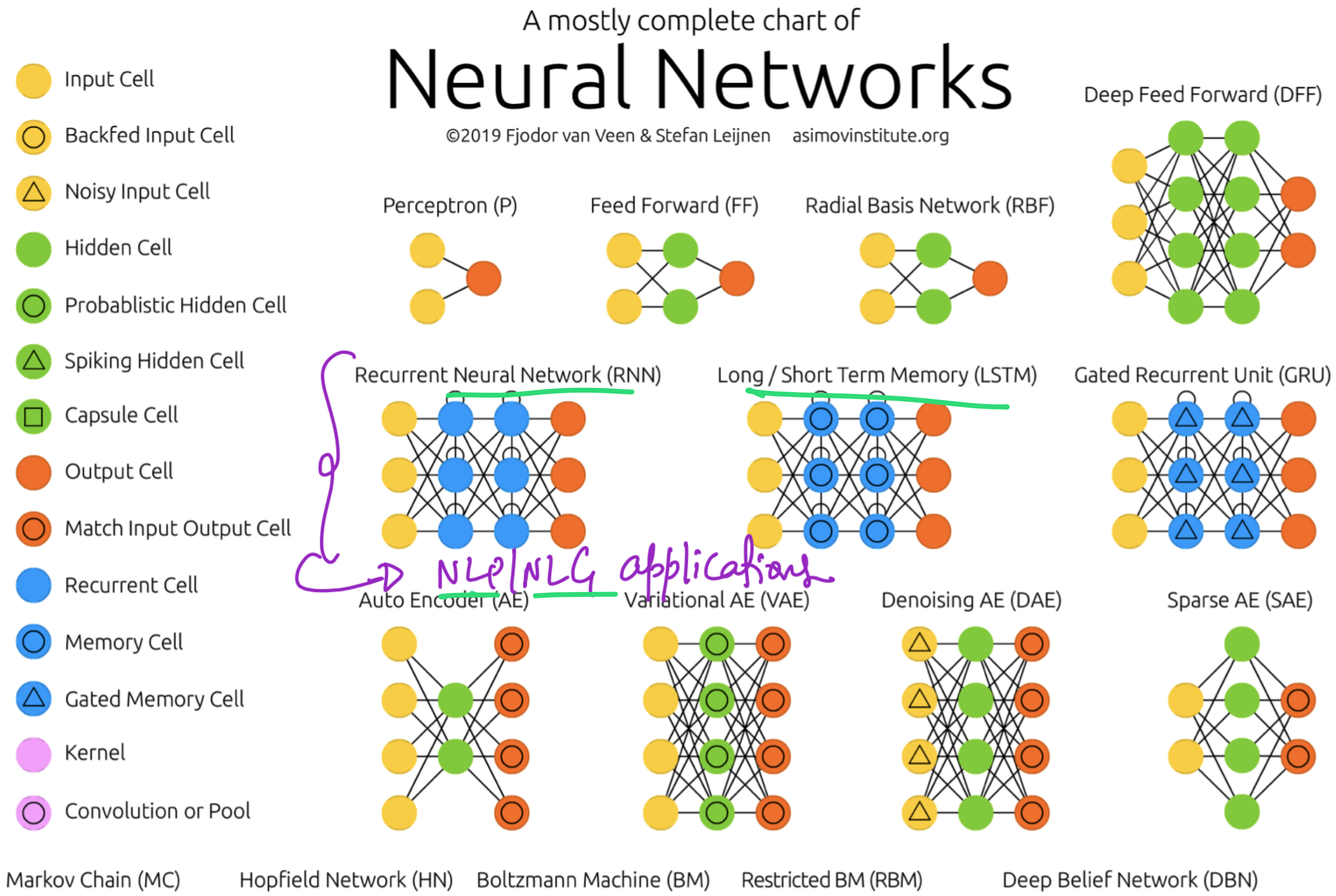
Deep Learning Zoo

Neural Networks Zoo



Deep Learning Zoo

Neural Networks Zoo

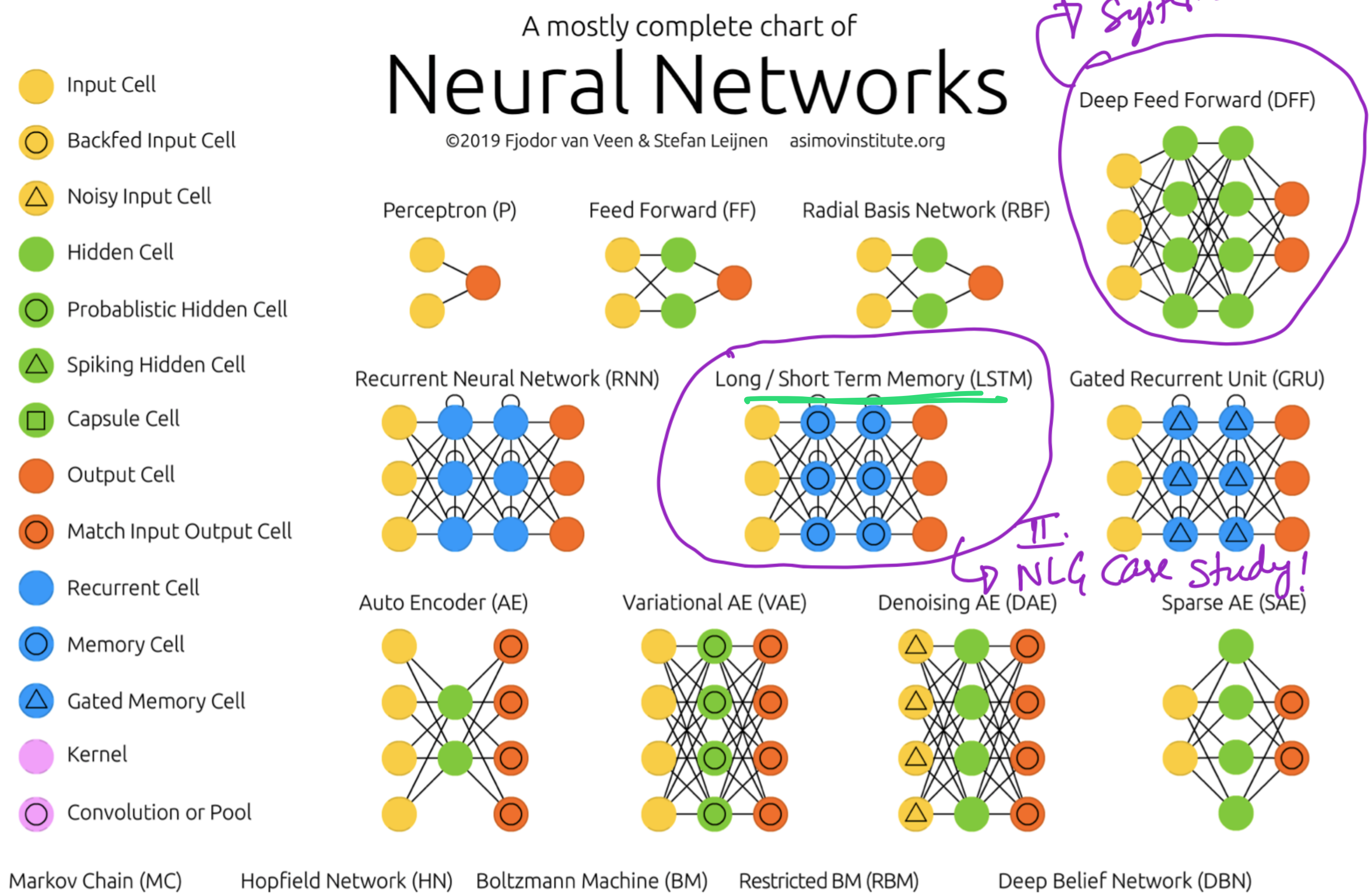


Deep Learning Zoo

Neural Networks Zoo

TODAY! I & II

I. Recommender Systems Case-Study

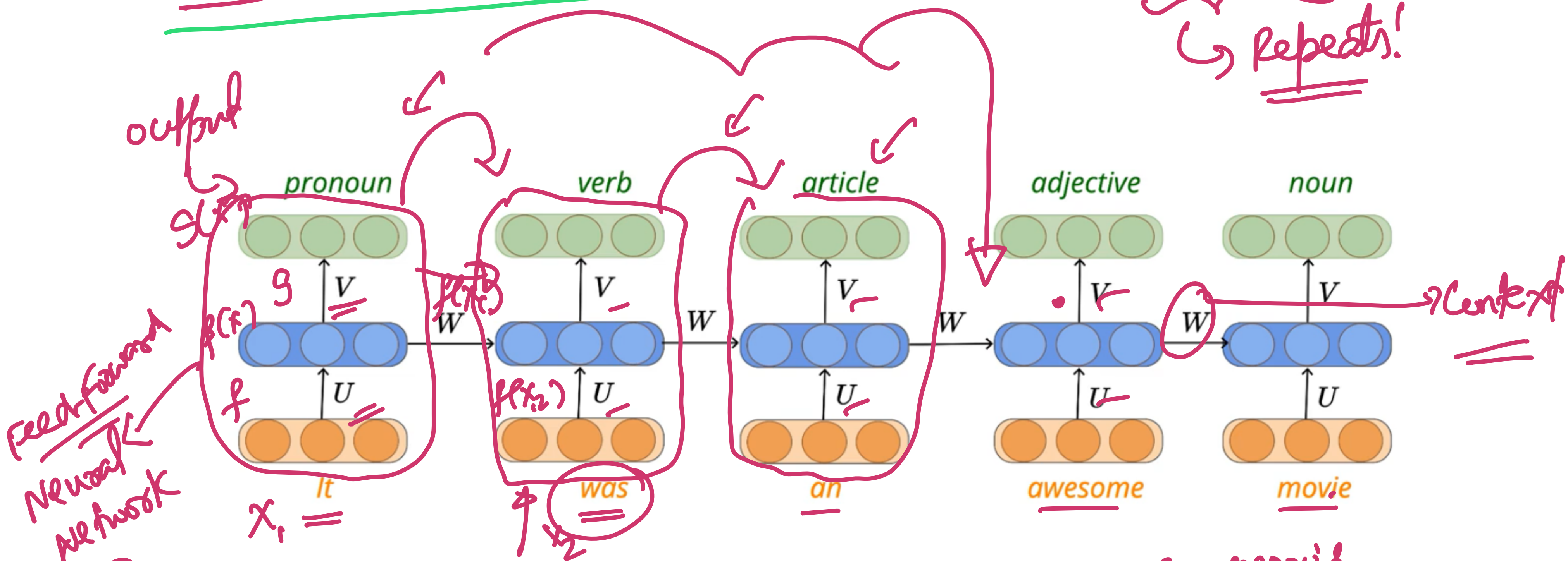


II. NLG Case Study!

LSTM

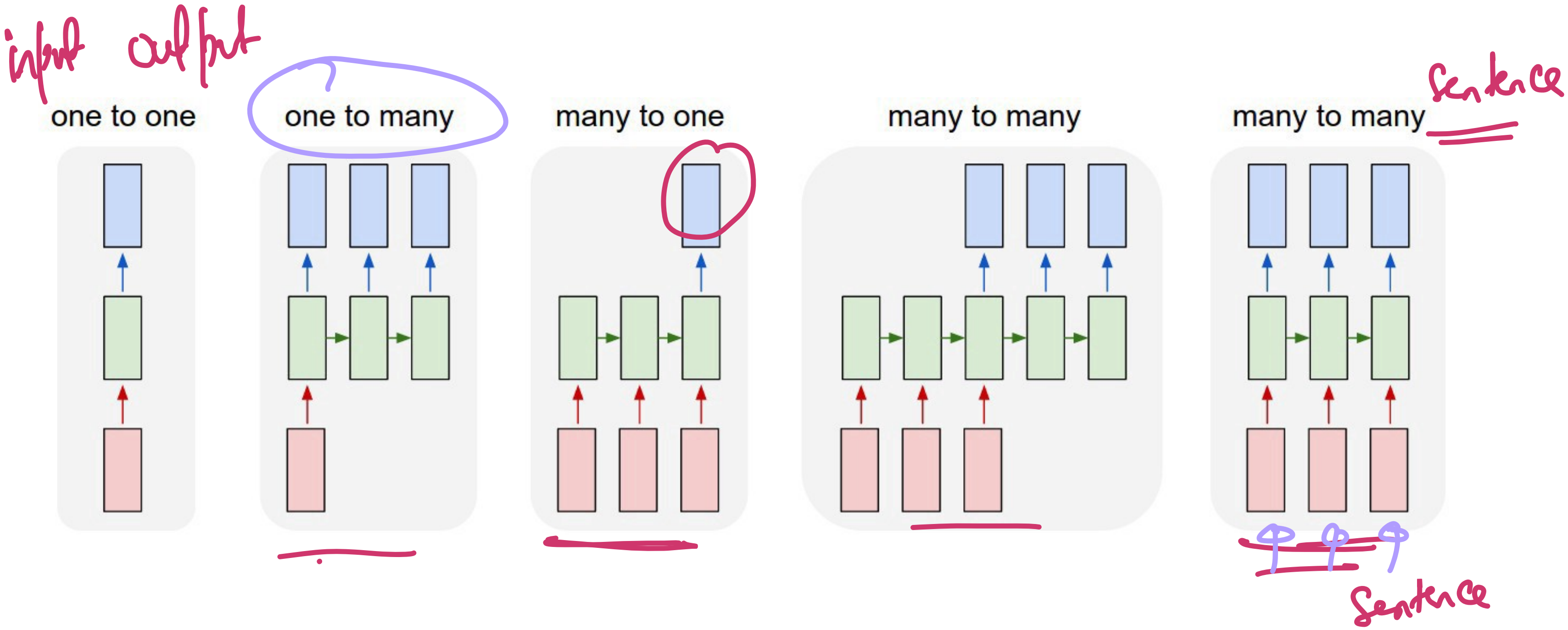
↳ Long Short Term Memory Network

RNN: Recurrent Neural network
↳ Repeats!



Input: It was an awesome movie

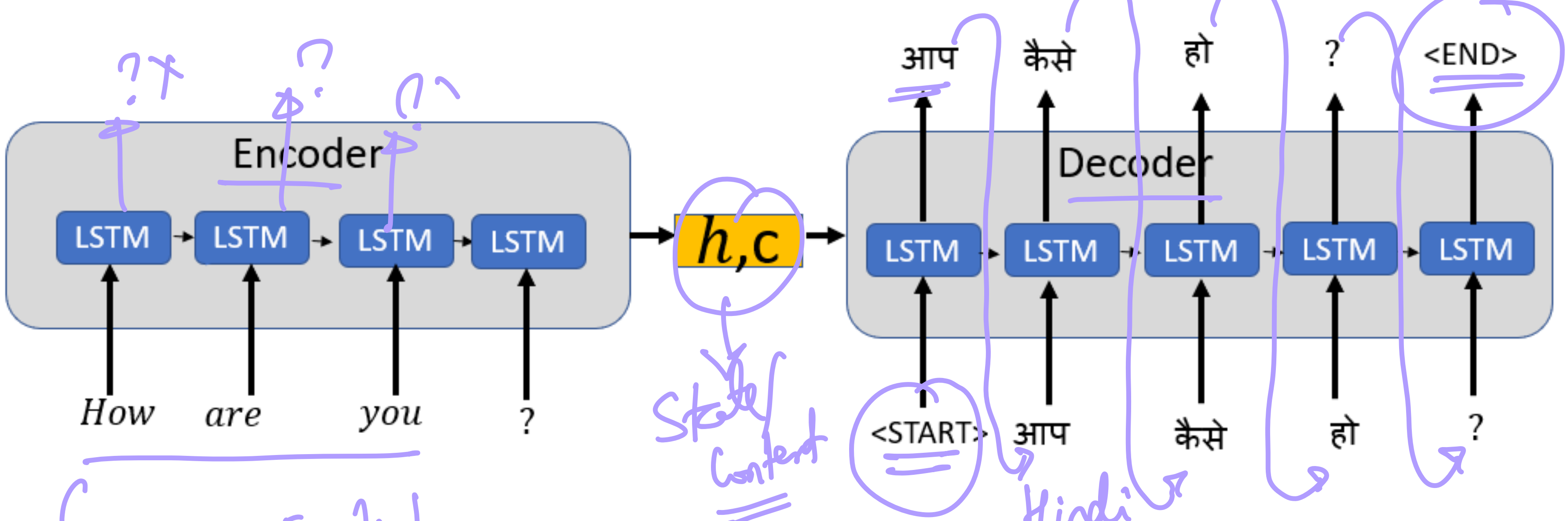
LSTM



input output
one to one
one to many
many to one
many to many
many to many
Sentence
Sentence

LSTM

Machine Translation



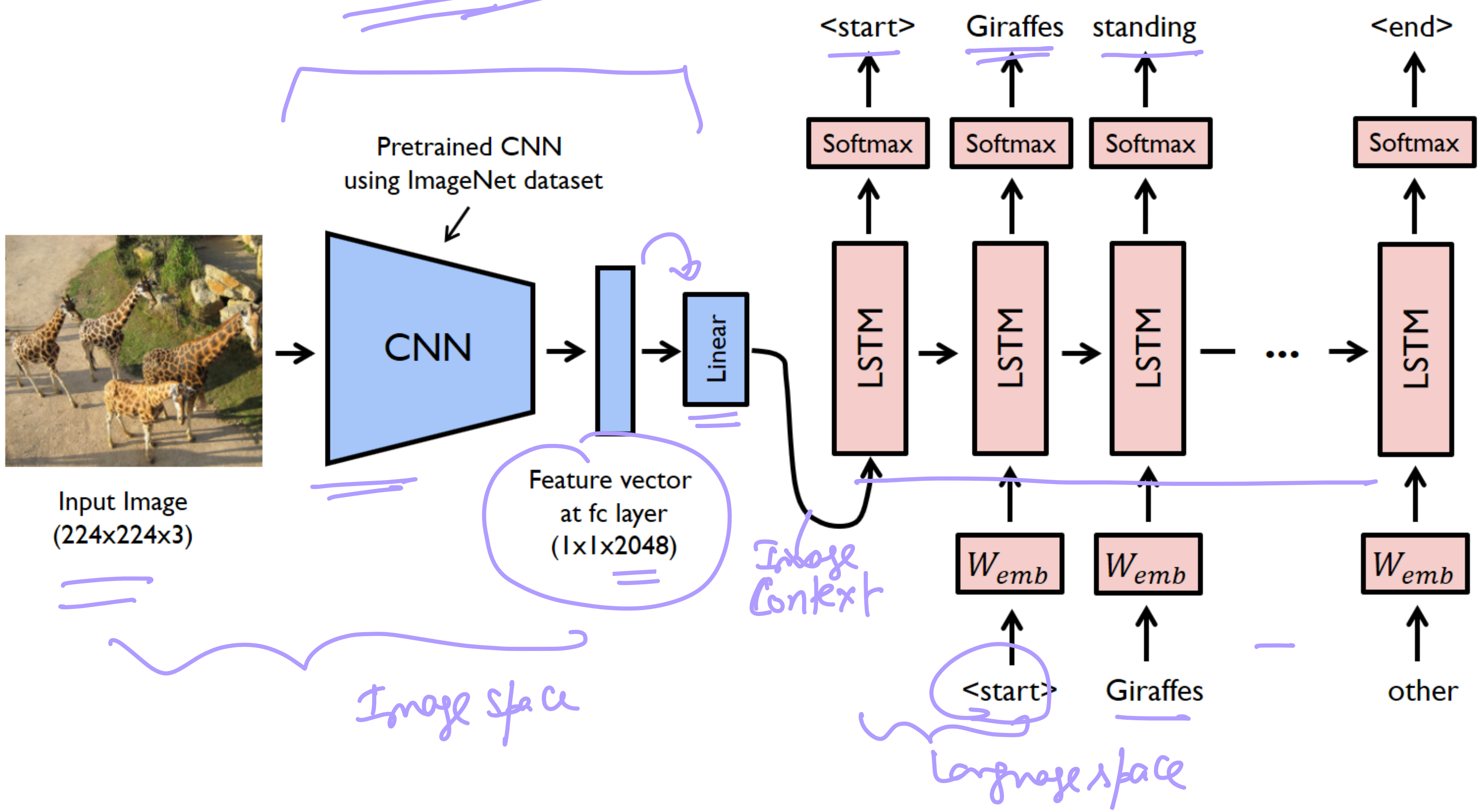
English

Encoding Input

LSTM

"Sequence Models"

Sequential token generation!



Metrics for Image Captioning

1. BLEU (Bilingual evaluation understudy)

Metrics for Image Captioning

1. BLEU (Bilingual evaluation understudy)
2. **METEOR** (Metric for Evaluation of Translation with Explicit Ordering)

Improvement on BLEU Score!

unigrams

Example:-

Pred:- Cows grazing on Grass

Truth:- Grazing Cows on grassy field

METEOR Metric

Pred:- Cows Grazing on Grass

Truth:- Grazing Cows on Grassy Field

$c = 3$

$\mu = 4$

1) Compute Precision:- (P) 3/4

2) Compute Recall:- (R) 3/5

3) F-score:- $\frac{10PR}{R+9P} = \frac{P \times R}{\frac{9}{10}P + \frac{R}{10}}$

↳ Biased towards Recall!

4) Fragmentation:-

$frag = \frac{c}{\mu}$

c: # chunks

μ: # tokens of output

5) Penalty

$p = 0.5 \times (frag)^3$

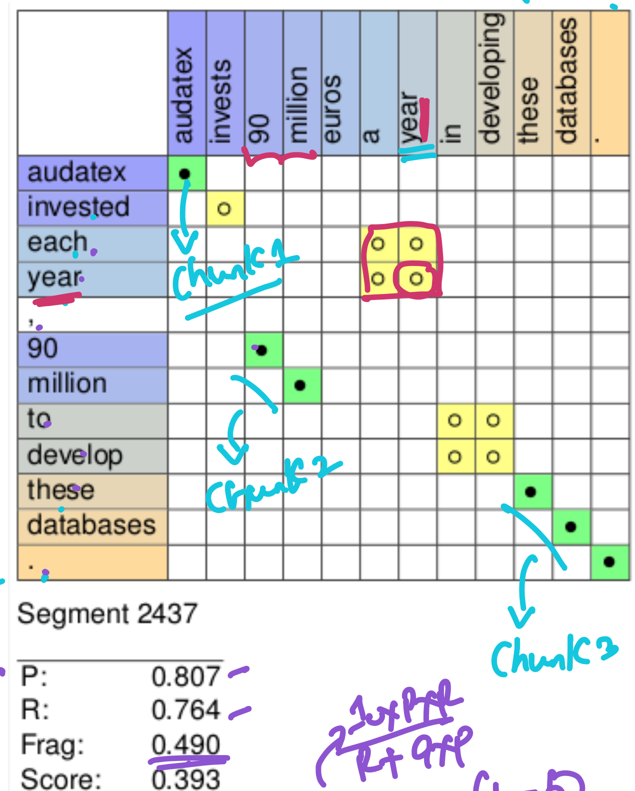
6) METEOR = $f\text{-}score\,(1-p)$

METEOR Example

Taken from Webpage on METEOR

Truth

Prediction



Chunk 1

Chunk 2

Chunk 3

Matrix

Include year

Frag $= \frac{c}{\mu} = \frac{4}{7} = 0.58$

Don't Include year

$= \frac{3}{6} = 0.5$

$\frac{6}{9}$

$\frac{10 \times P \times R}{R + 9 \times P}$

Meteor = F-Score × (1 − P) ; P = $0.5 \times \text{frag}^3$

ICE #2

$P = 5/6$

$R = 5/6$

F-score $= 5/6$

frag $= 3/5 = 0.6$ ✓

$p = 0.5 \times 0.6^3 = 0.108$

METEOR $= 5/6 \times (1-p) = 0.74$

Cows Grazing

Consider that your image captioning model generated the sentence: "Found Grazing cows on the grass" and the true caption was "Grazing cows found over the grass". What is the fragmentation value and METEOR score in this case?

Pred

Truth

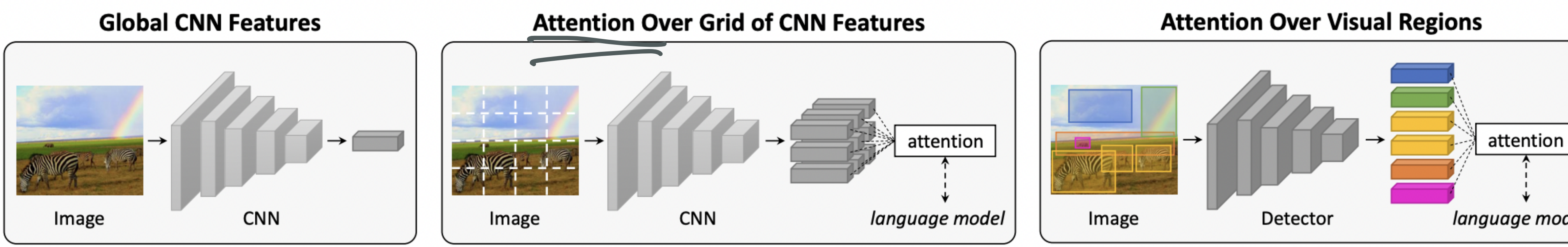
$\text{METEOR} = \text{F-Score} \times \underbrace{(1-p)}_{\text{penalty}}$

$\text{F-Score} = \dfrac{10 \times P \times R}{R + 9 \times P}$

$p = 0.5 \times (\text{frag})^3$

$\text{frag} = \dfrac{c}{u}$, $c \rightarrow$ chunks, $u \rightarrow$ # common words

Image Captioning Models



(a) (b) (c)

Fig. 2: Three of the most relevant visual encoding strategies for image captioning: **(a)** global CNN features; **(b)** fi grained features extracted from the activation of a convolutional layer, together with an attention mechanism guided the language model; **(c)** image region features coming from a detector, together with an attention mechanism.

Show&Tell

Breakout for Takeaways!

Discuss Takeaways (5 mins)

From today's lecture in your zoom group

Next Lecture

1. More on Image Captioning Models
2. Show and Tell Image Captioning Models